Workflow Systems and Frameworks: where are we headed?

Jim, Parag

October 2016

Why are we here?

- Start to do some brainstorming about software infrastructure 5-8 years out, covering
  - Frameworks
  - Workload and job management
  - Workflow orchestration
  - Data management
- Major questions to think about
  - What do we want our software products portfolio to look like?
  - What do we expect the world of computing to look like?
  - How do all these infrastructure products fit together? Should they?
  - Should we have a roadmap that addresses all these areas?
  - How does changes in technology affect us?
  - What R&D is necessary?
- The path towards the future
  - Funding: Who will pay for development? What can we afford to claim leadership in?
  - Partners: Who do we need to work with?
  - Timing: What needs to be done now? How much effort will it take? How much effort can be applied?
- Opportunity for two ends of the processing chain to meet

Background

- Laboratory core competency: Fermilab now has a core capability of advanced computer science, visualization and data
- Some Computational Science Theme laboratory objectives
  - GL-00510: Modernize scientific processes and access to computing
  - GL-00500: Continuously improve physics and infrastructure
  - GL-00530: Improve the partnership with the DOE Advanced Scientific Computing Research (ASCR) division and help drive the National Strategic Computing
  - GL-00520: Perform R&D to facilitate adoption of industry standards and emerging technologies
- Important initiative from Washington
  - National Strategic Computing Initiative (NSCI)
  - "The NSCI envisions: … A larger and more skilled HPC workforce that can take advantage of emerging technologies, including capabilities to support massive-concurrency, **data-intensive workflows**, tightly-coupled applications, and time-critical responses …"

http://computational-rd.fnal.gov/why-do-computational-science/

Software infrastructure of interest

Scientific Applications	Workload Management	Data Management	Workflow Management
• Frameworks • CMSSW • Art • ROOT?	• Jobsub • GlideinWMS • HTCondor	• SAM • IFDH • PhEDex • FTS • DBS	• Campaign management • Orchestration • WMAgent? • POMS?

- **Scientific Applications**: Experiment software/framework responsible for processing scientific data
- **Workflow Management**: Responsible for management and orchestration of computing campaigns and splitting the campaign into one or more computational jobs
- **Workload Management**: Responsible for queueing and managing individual computational jobs, acquiring resources to run them and managing these jobs and resources
- **Data Management**: Provides data cataloging, data management, data storage (archival?) and data movement

More questions to think about …

- How do we fit into the world of science?
  - High Throughput Computing (HTC),
  - Data Intensive Computing, Data Analytics
- How do we fit into the world of computing infrastructure?
  - Exascale, GRID
  - Pilot-based systems
- Collaborators and competitors
  - What about Pegasus? Are there other workflow systems?
  - What about HTCondor?
  - What about PanDA?
  - What about tools for distributed computing? (MPI, HPX, etc.)
- Where might our development funding come from?
  - ASCR, DOE Comp HEP and other R&D programs (CCE)
- CWP – we are participating. Our goals needs to be understood first, and then communicated through the white paper.

Funding Opportunities: Exascale Computing Project (ECP)

ECP goals will be tracked and accomplished via familiar DOE processes

- ECP will fund and manage work at the national laboratories, industry (including medium and small businesses), and universities
- In most cases ECP will provide incremental funding to teams that already have a funding base
  - Build on existing activities
  - "incremental" does not mean small
- There is a formal solicitation and selection process
- There are major deliverables and various reviews of major milestones and deliverables

8 Exascale Computing Project

ECP Holistic Structure

Capable exascale computing requires close coupling and coordination of key development and technology R&D areas.

ECP EXASCALE COMPUTING PROJECT

ECP Technical Approach

ECP will pursue a ten-year plan structured into four focus areas:

- **Application Development** deliver scalable science and mission performance on a suite of ECP applications that are ready for efficient execution on the ECP exascale systems.
- **Software Technology** enhance the software stack that DOE SC and NNSA applications rely on to meet the needs of exascale applications and evolve it to utilize efficiently exascale systems. Conduct R&D on tools and methods that enhance productivity and facilitate portability.
- **Hardware Technology** fund supercomputer vendors to do the research and development of hardware-architecture designs needed to build and support the exascale systems.
- **Exascale Systems** fund testbeds, advanced system engineering development (NRE) by the vendors, incremental site preparation, and cost of system expansion needed to acquire capable exascale systems.

9 Exascale Computing Project

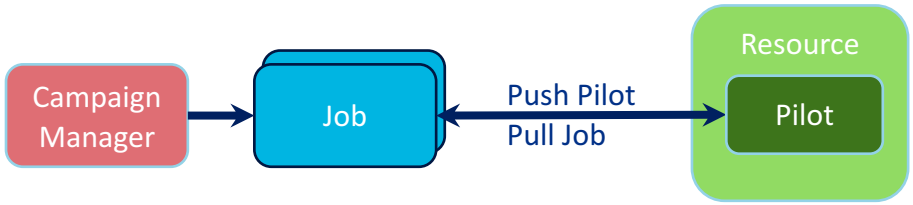
ECP EXASCALE COMPUTING PROJECT

Something else to consider …

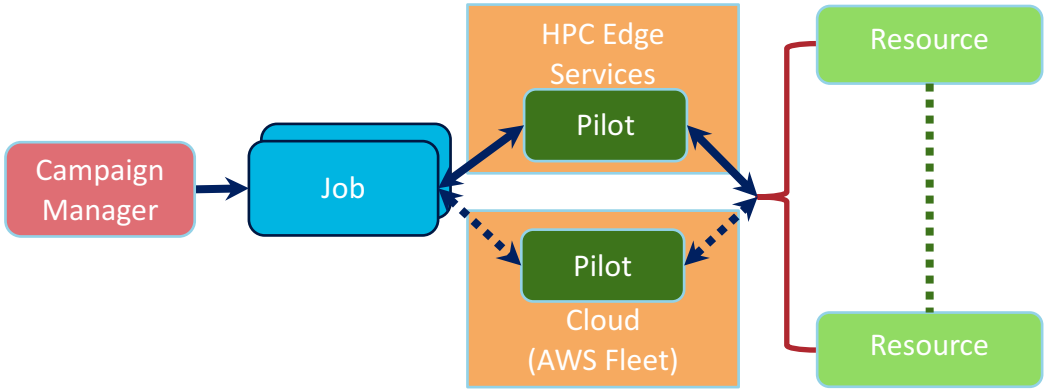
- **HEP Cloud**: Elastic Computing Facilities
- Focus on Computing Facilities
  - Supports multiple experiments and communities
- Elasticity to cope up with
  - Peak & Steady State Demand
  - Availability of different class of resources
    - Allocations for HPC
    - Budget for Cloud
    - Local and Opportunistic resources

WMS Stack & HPC Facilities

- WMS stack needs to adapt to new era of HPC and cloud fleet/auto-scaling resources
  - Acquire and manage a group of resources in one request
- There needs to be a more direct communication between the Campaign Management layer, Provisioning layer and the SWF
- Questions
  - Do we want HPC resources to look like HTC resources? How do the experiments want to utilize these resources?
  - Do we need to think about edge services and their functionalities?
  - Who are our potential collaborators?
  - What is our source of funding?



Pilot based systems today



Pilot based systems with edge services

Fermilab

Current state

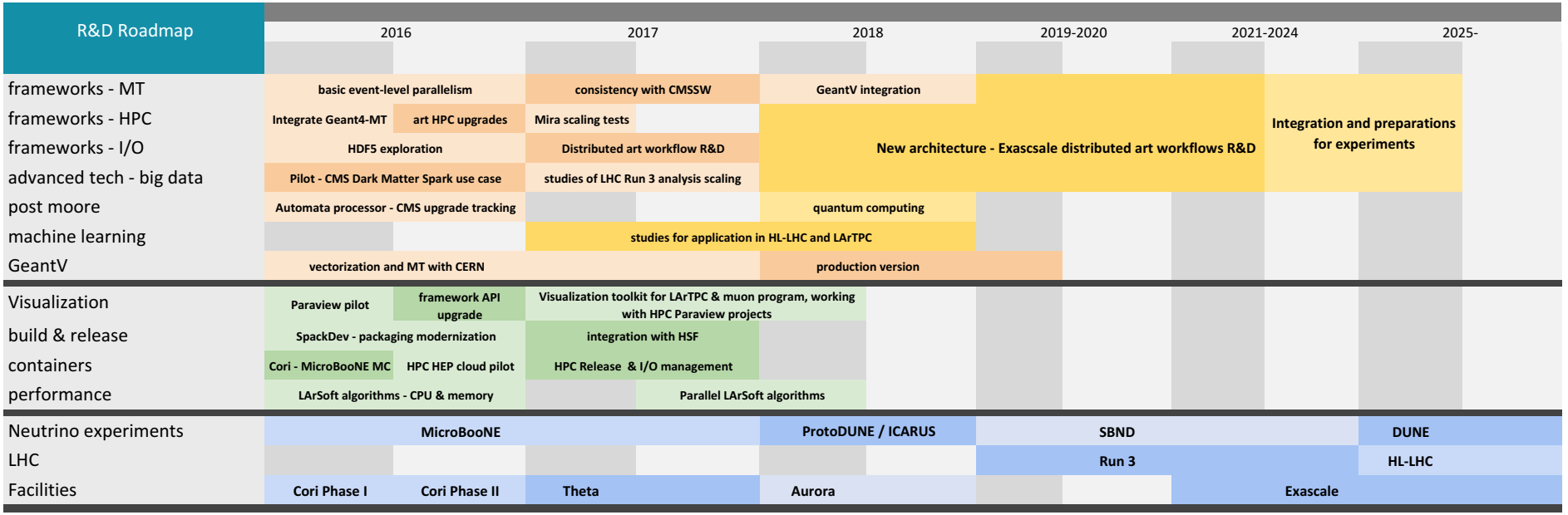
- Put in exascale proposals and white papers …
  - Two co-design centers (early on, a few years back)
  - SSIO & Visualization for data analytics (related topic, summer 2015)
  - A vision for integrated simulations, data reduction, and analysis (white paper)
  - Simulations & controls co-design center (summer 2016)
- Attended ASCR extreme-scale workflow workshop (4/2015)
  - Workflows white paper for exascale (cd-docdb 5551)
  - Both HTC and HPC present
  - *Distributed Area* (DA) and *Insitu* (IS) workflow management systems
- Got attention from ASCR leadership at operations review in May 2016
  - Wrote framework principals for ECP white paper
- Participated in the Goal Oriented Provisioning & Acquisition proposal in July 2016
- Workflow components and support libraries and services are starting to be defined and funded by ECP!
  - We did not participate in the first round of workflow and tools proposals, which will be getting funding soon. (reviews are complete)
  - Next round of calls for ECP software will likely be in late spring 2017
  - We need to think about this now.

Other support material

More information …

- Today, there is a fairly large barrier between the software framework orchestration layer and the higher-level campaign orchestration layer. This division limits efficient utilization of available resources.
- Over the next decade CMS will be realizing a pile up increase of 5-7 times. As the technology and its efficiency hardens, even a pile up increase of up to 8-10 times is a possibility. As the data rates increase, this will pose significant challenges in the campaign orchestration and software frame layer. There will be a need for both these layers to work together through to efficiently utilize the resources.
- HPC resources are soon becoming new key players. Current programming models that were developed for batch and computational grids and later adopted for Clouds, need to evolve to accommodate HPC with special focus on the availability of high performance interacts, many core compute resources and heterogenous architectures.
- Moving to increased data rates and to exa scale era machines will also mean evolving from file-oriented data access to a more efficient streaming-based data access. These changes to workflow & data access models will require rethinking the boundaries between campaign and SW framework layers.
- We need to understand how we as a lab want to position ourselves for the future and start planning on steps required to address these challenges.

R&D Roadmap



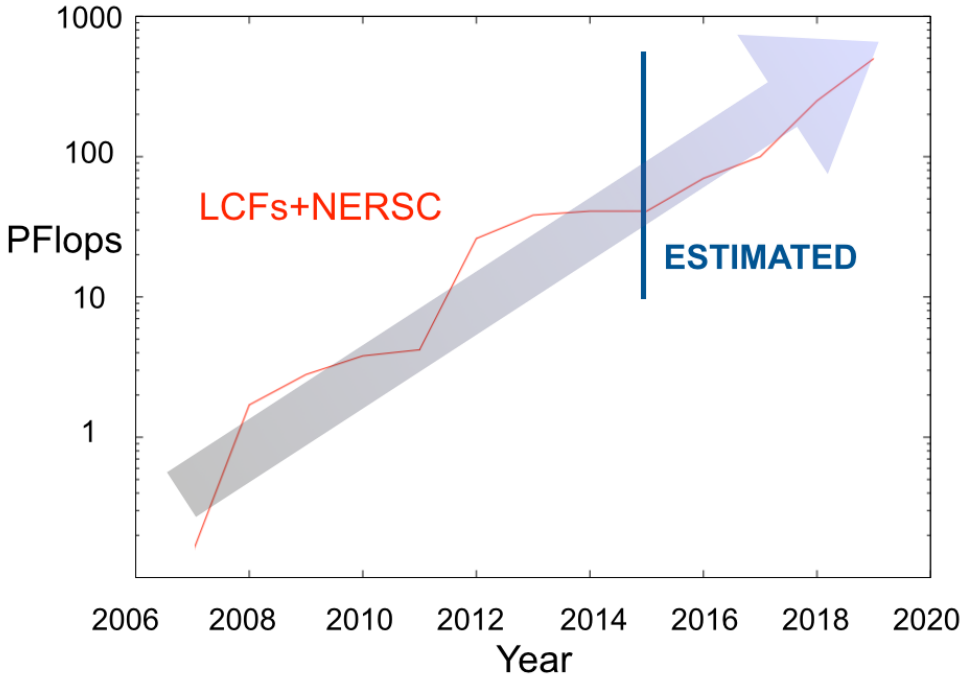
Forward looking | Meeting current needs | Transformative | Experiments & Facilities

- Addresses features that users need
- Includes partnerships with ASCR, national labs, CERN
- Matches experiment and facilities schedules
- Includes evolutionary and transformational changes

From FCE meeting: The Impact of HPC on HEP

- Not all problems can be solved using HPC systems, but many can (accelerators, cosmology, event generation/simulation, QCD…)
- Next generation of ASCR HPC machines (staging begins 2016, ends in 2018) will sum to ~500 petaflops of compute capability

- If HEP experiments use just 5% of that, i.e. 25 petaflops, it is ~25 times what the Grid will provide
- Learning how to leverage these resources to seamlessly supplement/enhance current capability is important
- New possibilities opened up by HPC platforms will offer unique computational opportunities



ASCR resources

ASCR Computing At a Glance

now ← → future

System attributes	NERSC Now	OLCF Now	ALCF Now	NERSC Upgrade	OLCF Upgrade	ALCF Upgrades	
Name Planned Installation	**Edison**	**TITAN**	**MIRA**	**Cori 2016**	**Summit 2017-2018**	**Theta**	**Aurora 2018-2019**
System peak (PF)	2.6	27	10	> 30	150	>8.5	180
Peak Power (MW)	2	9	4.8	< 3.7	10	1.7	13
Total system memory	357 TB	710TB	768TB	~1 PB DDR4 + High Bandwidth Memory (HBM) +1.5PB persistent memory	> 1.74 PB DDR4 + HBM + 2.8 PB persistent memory	>480 TB DDR4 + High Bandwidth Memory (HBM)	> 7 PB High Bandwidth On-Package Memory Local Memory and Persistent Memory
Node performance (TF)	0.460	1.452	0.204	> 3	> 40	> 3	> 17 times Mira
Node processors	Intel Ivy Bridge	AMD Opteron Nvidia Kepler	64-bit PowerPC A2	Intel Knights Landing many core CPUs Intel Haswell CPU in data partition	Multiple IBM Power9 CPUs & multiple Nvidia Voltas GPUS	2nd gen Intel Xeon Phi processor (code name Knights Landing)	3rd gen Intel Xeon Phi processor (code name Knights Hill)
System size (nodes)	5,600 nodes	18,688 nodes	49,152	9,300 nodes 1,900 nodes in data partition	~3,500 nodes	>2,500 nodes	>50,000 nodes
System Interconnect	Aries	Gemini	5D Torus	Aries	Dual Rail EDR-IB	Aries	2nd Generation Intel Omni-Path Architecture
File System	7.6 PB 168 GB/ s, Lustre®	32 PB 1 TB/s, Lustre®	26 PB 300 GB/s GPFS™	28 PB 744 GB/s Lustre®	120 PB 1 TB/s GPFS™	10PB, 210 GB/s Lustre initial	150 PB 1 TB/s Lustre®

U.S. DEPARTMENT OF ENERGY | Office of Science

LANS/LLNS May 29, 2015

BERKELEY LAB

WHAT'S EXASCALE LOOK LIKE?
(HYBRID CPU/GPU PATH)

Date	2009	2012	2017	CORAL-2 2023
System	Jaguar	Titan	Summit	Exascale
System peak	2.3 Peta	27 Peta	150+ Peta	1-2 Exa
System memory	0.3 PB	0.7 PB	2-5 PB	10-20 PB
NVM per node	none	none	800 GB	~2 TB
Storage	15 PB	32 PB	120 PB	~300 PB
MTTI	days	days	days	O(1 day)
Power	7 MW	9 MW	10 MW	~20 MW
Node architecture	CPU 12 core	CPU + GPU	X CPU + Y GPU	X loc + Y toc
System size (nodes)	18,700	18,700	3,400	How fat?
Node performance	125 GF	1.5 TF	40 TF	depends (X,Y)
Node memory BW	25 GB/s	25 - 200 GB/s	100 – 1000 GB/s	10x fast vs slow
Interconnect BW	1.5 GB/s	6.4 GB/s	25 GB/s	4x each gen
IO Bandwidth	0.2 TB/s	1 TB/s	1 TB/s	flat

Al Geist: NITRD talk 4/16/2015

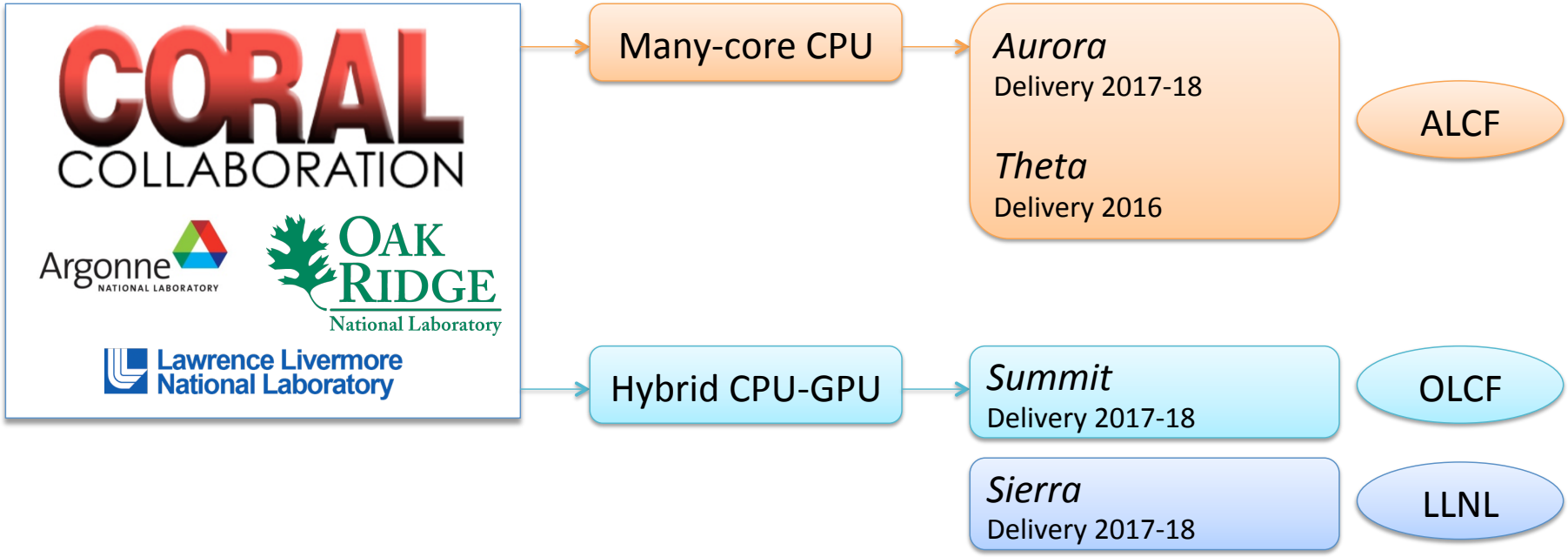
WHAT'S EXASCALE LOOK LIKE?
(MANY-CORE PATH)

				CORAL-2
Date	**2008**	**2012**	**2018**	**2023**
System	**Intrepid**	**Sequoia**	**Aurora**	**Exascale**
System peak	0.6 Peta	20 Peta	180+ Peta	**1-2 Exa**
System memory	0.08 PB	1.6 PB	>7 PB	**10-50 PB**
MTTI	weeks	weeks*	days	**O(1 day – 1 week)**
Peak Power	2 MW	9.6 MW	~13 MW	**~25 MW**
Node architecture	CPU 4 core	CPU 16 core	CPU >72 core	**CPU + ?**
System size (nodes)	40,960	98,304	>50,000	**~100K**
Node performance	13.6 GF	204.8 GF	-	**O(10 TF)**
Node memory BW	13.6 GB/s	42.5 GB/s	-	**50x fast vs slow**
Interconnect BW	5.1 GB/s	40 GB/s	-	**4x each gen**
Storage	6 PB	50 PB	>150 PB	**~1000 PB**
IO Bandwidth	80 GB/s	1TB/s	>1 TB/s	**flat**

* Mira MTTI

HPC systems

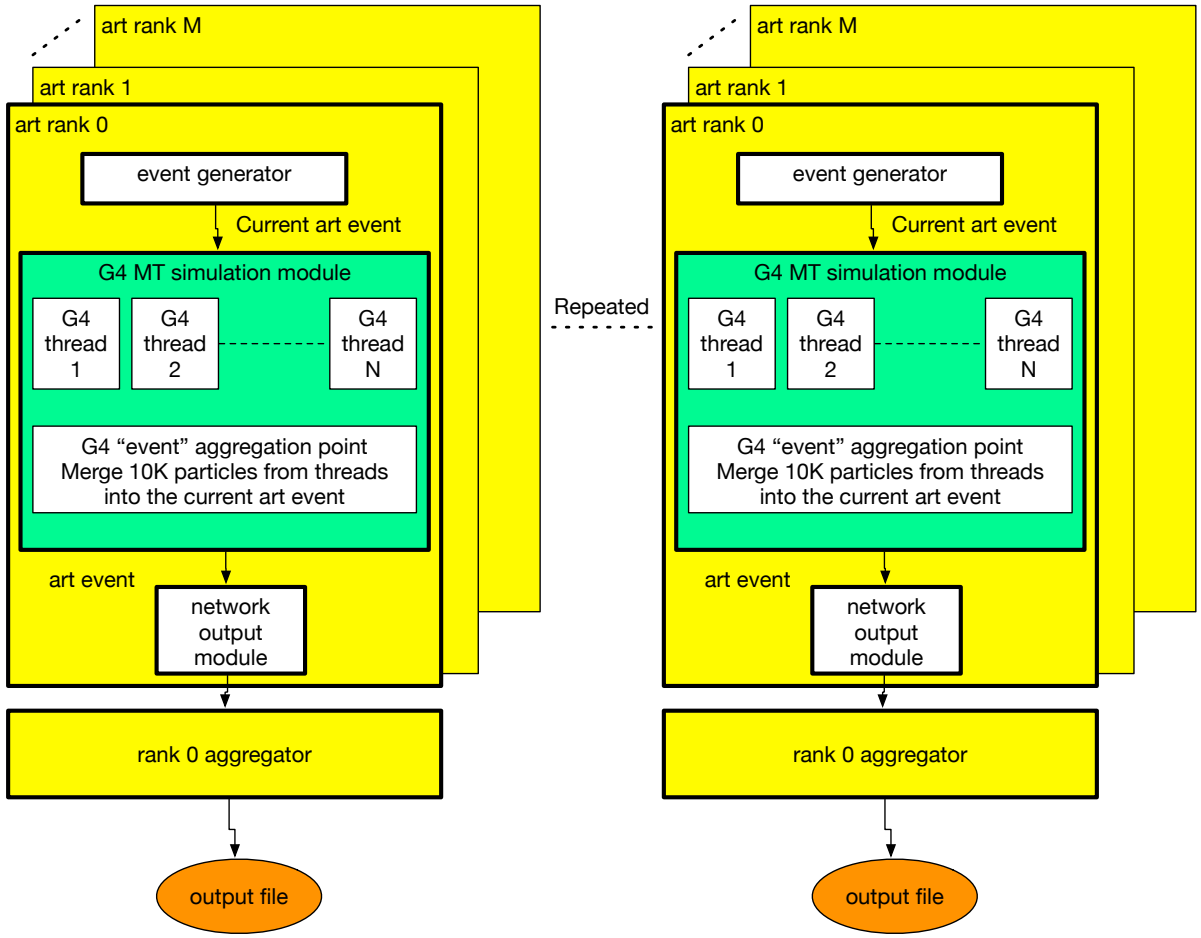
Next-Generation ALCF-3 System



Aurora, Summit are LCF's **pre-exascale** systems.

art-HPC

- **Extending the ART Framework to Support Large Scale Multiprocessing for the Intensity Frontier**
  - Partnership with Tom LeCompte at ANL
  - Migration of art to HPC and Mira
  - Using MPI
  - Multi-threaded Geant4
- Target is to produce 10^12 muons for muon g-2 on ALCF Mira
- Architected to addressed
  - limit I/O to filesystem
  - scaling



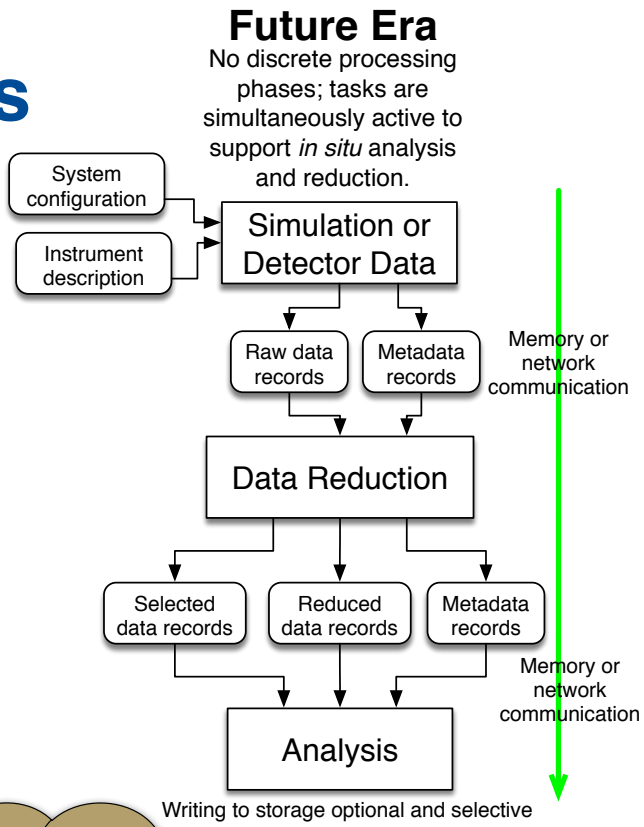
NOTE: Same architecture applied to running a multi-parameter tuning of event generators using collider data analysis on Mira using Pythia

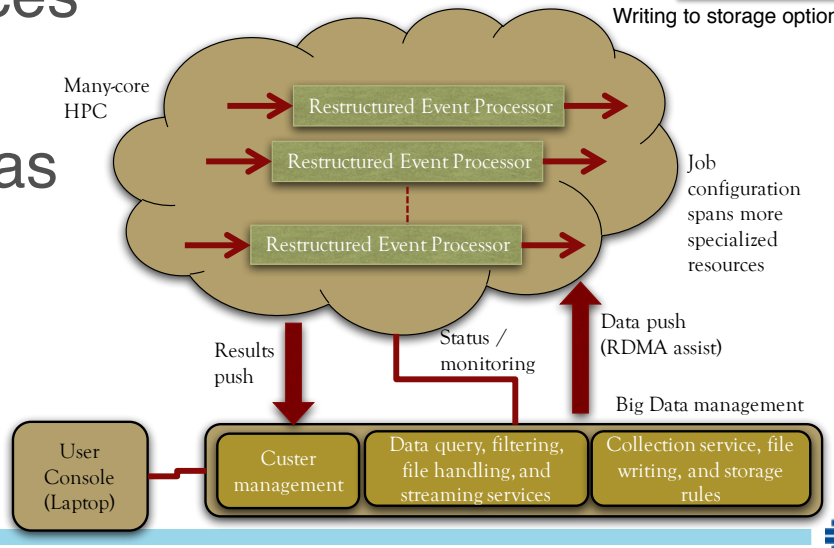
https://cdcvs.fnal.gov/redmine/projects/art-hpc/wiki/

Fermilab

Evolving architecture key elements

- Limit I/O to disk storage
- Utilize large vector units
- Efficient movement of data between processes in a distributed environment using high bandwidth networking
- Shared framework services within nodes
- Localized data caching, as in big data technology
- Tighter integration with workload / workflow management





Fermilab

More from NCIS

- Developing public-private collaboration to ensure broad deployment of NSCI-developed capabilities is a common thread among the NSCI objectives. Architectures and software systems that increase coherence between computational and data intensive workflows will facilitate the convergence of modeling and data analytics. This, in turn, may provide more capabilities to enable the business and scientific enterprises of the future. Breaking through the limitations of Moore's Law is imperative to producing compact and power-efficient systems, bringing current-day HPC capabilities to new sectors, and establishing new frontiers in computing and analytics.
- Objective 2: The NSCI seeks to develop a coherent platform for modeling, simulation, and data analytics, primarily through the development of a more agile and reusable HPC software portfolio. Historically, there has been a separation between data analytic computing and modeling and simulation. Systems have been optimized for a specific class of applications, but the differences between these application spaces are fading rapidly. The growth of extremely large-scale data analytics within the modeling and simulation community demands a dynamic interaction between analysis and simulations. As data analytics increases in computational intensity, and modeling and simulation encounter increasingly complex problems, both fields face barriers to scalability along with new demands for interoperability, robustness, and reliability of results.
- Box 3: Convergence of Data Analytic Computing and Modeling and Simulation. Historically, there has been a separation between data analytic computing and modeling and simulation. Data analytics focuses on inferring new information from what is already known to enable action on that information. Modeling and simulation focuses on insights into the interaction of the parts of a system, and the system as a whole, to advance understanding in science and engineering and inform policy and economic decision-making. While these systems have traditionally relied on different hardware and software stacks, many of the current challenges facing the two disciplines are similar. The growth of extremely large-scale data analytics within the modeling and simulation community demands a dynamic interaction between analysis and simulations. As demands for computational intensity increase, the data analytics community faces barriers to scalability along with new demands for interoperability, robustness, and reliability of results. A coherent platform for modeling, simulation, and data analytics would benefit both disciplines while maximizing returns on R&D investments. The primary challenges lie within and across software layers of the HPC environment. In particular, a more agile and reusable HPC software portfolio that is equally capable in data analytics and modeling and simulation will improve productivity, increase reliability and trustworthiness in computations, and establish more sustainable yet agile software. These improvements, in turn, may provide mutual benefit for the analytics and simulation communities, along with new advances across industry, academia, and government.